Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 54 (2012) 459–476

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence

Jordi Adell [a,*], David Escudero [b], Antonio Bonafonte [a]

[a] *Universitat Politècnica de Catalunya, Barcelona, Spain*
[b] *Universidad de Valladolid, Valladolid, Spain*

Received 14 March 2011; received in revised form 28 October 2011; accepted 29 October 2011
Available online 15 November 2011

Abstract

Until now, speech synthesis has mainly involved reading-style speech. Today, however, text-to-speech systems must provide a variety of styles because users expect these interfaces to do more than just read information. If synthetic voices must be integrated into future technology, they must simulate the way people talk instead of the way people read. Existing knowledge about how disfluencies occur has made it possible to propose a general framework for synthesising disfluencies. We propose a model based on the definition of disfluency and the concept of underlying fluent sentences. The model incorporates the parameters of standard prosodic models for fluent speech with local modifications of prosodic parameters near the interruption point. The constituents of the local models for filled pauses are derived from the analysis corpus, and constituent's prosodic parameters are predicted via linear regression analysis. We also discuss the implementation details of the model when used in a real speech synthesis system. Objective and perceptual evaluations showed that the proposed models outperformed the baseline model. Perceptual evaluations of the system showed that it is possible to synthesise filled pauses without decreasing the overall naturalness of the system, and users stated that the speech produced is even more natural than the one produced without filled pauses.

Keywords: Speech synthesis; Conversational speech; Talking speech synthesiser; Filled pause; Disfluency; Underlying fluent sentence; Prosody; Ogmios; Perceptual evaluation

1. Introduction

Speech synthesis has already reached a high standard of naturalness (Aaron et al., 2005), mainly due to the use of effective techniques such as unit selection (van Santen et al., 1997; Taylor, 2009) and other new technologies based on Markov models (Zen et al., 2009) that are increasing in popularity. Until now, the main product of speech synthesis has been reading-style speech; it was felt that this style could be used in any other situation. Today, however, the various applications of text-to-speech (TTS) systems (e.g., video-games, robotics, dialogue systems, audiovisual production, multilingual broadcasting and automatic film dubbing) demand a variety of styles. Systems must be more expressive because users expect interfaces to do more than just read information (Lee, 2010).

1.1. Motivation

If synthetic voices are to be integrated into future technology, they must simulate the way people talk instead of the way people read. TTS systems will have to be able to generate speech suitable for conversations. Therefore, we argue that it is necessary to transition from *reading-style* to *talking-style* speech synthesisers. The two styles differ significantly due to the inclusion of a variety of linguistic resources. Disfluencies are the ones for handling the time course of speech production. They are defined as phenomena that interrupt the flow of speech and do not add

[*] Corresponding author. Tel.: +34 931939046.
E-mail address: jordi.adell@upc.edu (J. Adell).

0167-6393/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2011.10.010

propositional content to an utterance (Fox Tree, 1995). However, despite their lack of propositional content, disfluencies are argued not to be problems in speaking but solutions to problems in speaking (Clark, 2002). Disfluencies have a communicative value; for instance, they facilitate vocalisation synchronisation between addressees in conversations (Clark, 2002), improve listening comprehension by creating a delay (Rose, 1998; Fox Tree, 2001) or indicate the complexity of the upcoming message (Watanabe et al., 2005). As a result, disfluencies have been observed to be a very frequent phenomenon in spontaneous speech (Tseng, 1999). Filled pauses are one of the most frequent type of disfluency and consequently is the most studied one. This paper presents a method of generating filled pauses in unit selection TTS systems.

The study of filled pauses has been approached not only from a phonetic perspective (Shriberg, 1999; Tseng, 1999) but also from the perspective of more technological fields, mainly because of the need to model this frequent phenomenon to improve spontaneous automatic speech recognition (Nakatani and Hirschberg, 1994; Shriberg et al., 1997). Also, in the context of dialogue systems, some authors report the use of disfluencies to classify a speaker's communicative intention (Savino and Refice, 2000). Furthermore, there are studies that have approached this problem from a psycholinguistic perspective, analysing the use of filled pauses to coordinate the speech actions of speakers and addressees (Clark, 2002) and to create delays (Clark and Fox Tree, 2002; O'Connell and Kowal, 2004). The fact that filled pauses serve a communicative function supports our claim that including filled pauses in synthetic speech is necessary to create highly natural synthetic voices for general purposes.Because humans use filled pauses, machines will potentially sound more humanlike if they do so. Furthermore, machines may also need to stop for processing reasons in incremental production, and rather than just pause, they could do something more humanlike by putting in a filled pause.

1.2. Scope

To appropriately use filled pauses in TTS systems requires one to take into account several linguistic considerations: one must select the situation in which the filler is used, identify the correct place in which to include it in the sentence and determine what the filler should sound like. Linguistics research is developing guidelines for predicting filled pauses based on empirical data (Andersson et al., 2010; Adell et al., 2007; Pakhomov, 1999; Stolcke and Shriberg, 1996). In this paper, we focus on what the filler pause should sound like. We feel that being able to generate filled pauses in a realistic manner is the first step towards inserting filled pauses with a specific meaning in the correct position within speech. We present a system that allows users to easily enter filled pauses into the speech stream, giving them the opportunity to potentially increase the system's communicative capabilities.

1.3. Background

The synthesis of disfluent speech within the framework of unit-selection speech synthesis presents a number of challenges. First of all, most existing unit selection systems have a closed inventory that does not contain disfluencies. Therefore, the state-of-the-art machine learning techniques, that are usually applied to model prosody using existing data are not useful here. Secondly, both prosodic models and text analysis models (e.g., part-of-speech tagging) expect well-structured sentences. When fluency is disrupted, the expected structure is also disrupted, making it more difficult for standard models to predict the necessary parameters.

During the last few years, there has been growing interest in the implementation of expressive speech synthesis systems. Eide (2002) and Hamza et al. (2004) considered the inclusion of paralinguistic events such as filled pauses in these systems reporting improvements in the naturalness of synthetic speech. In addition, Sundaram and Narayanan (2002, 2003) reported benefits of including filled pauses and other *VoiceFonts* (Campbell, 1998) in a limited domain synthesiser. More recently, Andersson et al. (2010) proposed a more ambitious approach based on the recording of a large spontaneous database. These approximations consider the filler as a unit in the speech database that is used to compose the final utterance. Nevertheless, in our previous phonetic studies of disfluency, we reported a local acoustic impact of disfluency on the surrounding syllables (Adell et al., 2008). As a result, based on the model proposed by Shriberg (1999), we propose an alternative solution, that considers both the potential fluent sentences associated with the disfluent sentence and the local modifications produced when the disfluency occurs. These local modifications can affect speech prosody and quality of the delivery. Taking this framework as a starting point, we applied it to the synthesis of filled pauses and validated its relevance for this specific type of disfluency. Because the results have been successful, we feel that it will be productive to work with other types of disfluencies in the future.

1.4. Overview

Our proposal is supported by Adell et al. (2010a,b), and here we present the fundamentals of the model, the use of speaker-dependent operative prosodic rules to generate synthetic fillers, the details of the TTS implementation process and the results of several perceptual tests that validate the proposed method.

In Section 2, we present a review of previous work on disfluent speech and the underlying fluent sentence model. The main components of the model are the use of appropriate fluent sentences to construct the disfluent one and prosodic local modifications around the interruption point. Therefore, standard models can be used to predict most